ABRF Proteome Informatics Research Group (iPRG) 2016 Study: Inferring Proteoforms from Bottom-up Proteomics Data

Joon-Yong Lee,[1] Hyungwon Choi,[2] Christopher M. Colangelo,[3] Darryl Davis,[4] Michael R. Hoopmann,[5] Lukas Käll,[6] Henry Lam,[7] Samuel H. Payne,[1] Yasset Perez-Riverol,[8] Matthew The,[6] Ryan Wilson,[1] Susan T. Weintraub,[9] and Magnus Palmblad[10,]*

[1]Pacific Northwest National Laboratory, Richland, Washington 99352, USA; [2]National University of Singapore, 117547 Singapore, Singapore; [3]Agilent Technologies, 121 Hartwell Ave., Lexington, MA 02421; [4]Janssen Research and Development, LLC, Spring House, Pennsylvania 19087, USA; [5]Institute for Systems Biology, Seattle, Washington 98109, USA; [6]Science for Life Laboratory, KTH - Royal Institute of Technology, 171 65 Solna, Sweden; [7]Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China; [8]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom; [9]Department of Biochemistry and Structural Biology, The University of Texas Health Science Center, San Antonio, Texas 78229, USA; and [10]Center for Proteomics and Metabolomics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands

This report presents the results from the 2016 Association of Biomolecular Resource Facilities Proteome Informatics Research Group (iPRG) study on proteoform inference and false discovery rate (FDR) estimation from bottom-up proteomics data. For this study, 3 replicate Q Exactive Orbitrap liquid chromatography-tandom mass spectrometry datasets were generated from each of 4 *Escherichia coli* samples spiked with different equimolar mixtures of small recombinant proteins selected to mimic pairs of homologous proteins. Participants were given raw data and a sequence file and asked to identify the proteins and provide estimates on the FDR at the proteoform level. As part of this study, we tested a new submission system with a format validator running on a virtual private server (VPS) and allowed methods to be provided as executable R Markdown or IPython Notebooks. The task was perceived as difficult, and only eight unique submissions were received, although those who participated did well with no one method performing best on all samples. However, none of the submissions included a complete Markdown or Notebook, even though examples were provided. Future iPRG studies need to be more successful in promoting and encouraging participation. The VPS and submission validator easily scale to much larger numbers of participants in these types of studies. The unique "ground-truth" dataset for proteoform identification generated for this study is now available to the research community, as are the server-side scripts for validating and managing submissions.

Key Words: inference, false discovery rate, community study, best practice

INTRODUCTION

In bottom-up proteomics, one begins by digesting the proteins in complex biologic samples into even more complex mixtures of peptides. These peptides, rather than the proteins, are the analytes that are identified and quantified based on one or more types of mass spectrometry (MS) data. The direct linkage between peptides and proteins becomes lost in the digestion step. Protein inference[1] is a significant challenge for the proteomics experiments and is often addressed with a combination of genomic sequences and statistical methods to infer information about the proteins from the measured peptides. Homologs—related and similar proteins—often share peptides. Some peptides are exclusively present in one protein, some are common to several members of a protein family, and yet others are shared among multiple protein families. Furthermore, through various post-transcriptional and post-translational processes, one gene can give rise to many proteoforms. At each step in the process [peptide-spectrum matching, unique peptide identification, protein (gene) identification, and proteoform inference], the corresponding false discovery rate (FDR) should be calculated and propagated to the next step. Strict and accurate FDR control is particularly important in the analysis and integration of large numbers (millions or even billions) of tandem MS (MS/MS). Recent studies

*Address correspondence to: Center for Proteomics and Metabolomics, Leiden University Medical Center, P.O. Box 9600, 2300 RC Leiden, The Netherlands. E-mail: n.m.palmblad@lumc.nl
doi: 10.7171/jbt.18-2902-003

have concluded that the lack of complexity in gold standard datasets limits their applicability to protein inference algorithm benchmarking.[2, 3]

In the 2016, Association of Biomolecular Resource Facilities (ABRF) Proteome Informatics Research Group (iPRG) study, "Inferring Proteoforms from Bottom-up Proteomics Data," participants were invited to evaluate methods for proteoform inference. To generate the dataset for the study, partially overlapping oligopeptides [protein epitope signature tags (PrESTs)] were expressed in *Escherichia coli* and spiked at differing proportions into four samples of a common background to mimic mixtures of protein homologs for proteoform identification. Participants were provided with a protein sequence database that contained the proteoforms that were present, along with a number of similar but absent proteoforms. Based on our experience from previous studies, especially the 2015 study,[4] we decided to develop and test a new submission mechanism, wherein a strict format validator gave direct feedback to the participants and only accepted correctly formatted submission files. The format validator, as well as the data, instructions, and all other study-related material, was hosted on a virtual private server (VPS), accessible at www.iprg2016.org. This arrangement was selected to minimize the time spent by iPRG members to verify manually and align the file formats that were submitted. Another novelty in 2016 was the ability to submit method descriptions as R Markdown[5, 6] or IPython (Jupyter) Notebooks[7, 8] for easy sharing and comparison of strategies for proteoform inference and FDR calculation. The new submission mechanics and submission formats were also evaluated as part of the study.

MATERIALS AND METHODS

MS and sequence data

PrESTs, 383 partially overlapping oligopeptides, were originally expressed in *E. coli* for the Human Protein Atlas Project[9] and graciously provided for this study as mimics for protein homologs to create "ground-truth" datasets for the analysis of proteoforms. Four samples were prepared by spiking different combinations of PrESTs into a common background (**Fig. 1**). The samples were analyzed by data-dependent liquid chromatography-MS/MS using higher-energy collisional dissociation on a Q Exactive MS (Thermo Fisher Scientific, Waltham, MA, USA), as described in detail by The *et al.*[10]

Participants were allowed to use their method of choice for peptide-spectrum matching, based on MS1 and/or MS2 data. A FASTA sequence database was provided for which participants were instructed to use without modification. The database contained the PrEST sequences, 5592 *E. coli* proteins, and 1000 sequences of PrESTs not present in any of the samples. The participants were asked to report only results on the PrESTs detected in the samples [named in the database as "HPRR," (Human Proteome Resource, Recombinant) followed by a unique number].

Submission and anonymization process

For the results report, participants were asked to provide a tab-delimited table listing the proteoforms identified in each of the 12 datasets (4 samples with 3 replicates of each), along with the probabilities of presence [labeled as posterior error probability (PEP)] or q values for the identifications ("FDR"). Each row of the data matrix started with the identified HPRR proteoform accession number, followed by 12 columns for the FDRs/PEPs for each sample (labeled by dataset name: A1, A2, . . ., D3). The first row was designated as a header row. It was not revealed to the participants that the "A" datasets were generated from the sample spiked with both PrEST pool A (192 PrESTs) and PrEST pool B (191 PrESTs), the "B" datasets corresponded to the B pool, the "C" datasets the A pool, and the "D" datasets the *E. coli* background by itself. The submitted

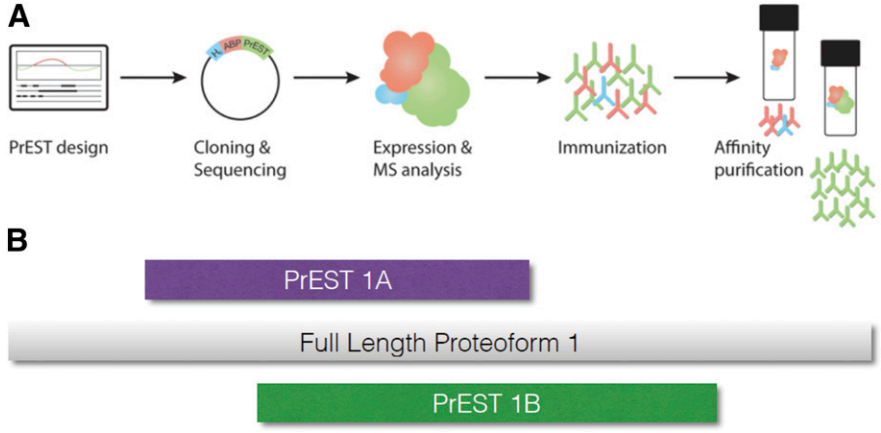

FIGURE 1

For the 2016 iPRG study, the dataset was generated from 4 samples of different combinations of small proteins (PrESTs), expressed in *E. coli* in an established pipeline for generating antibodies and spiked into a constant background of *E. coli* proteins (A). The PrESTs were overlapping by pairs (B), giving rise to some unique and some shared tryptic peptides for each PrEST.

spreadsheets were validated during submission by a Python script, specifically developed for this study. The instructions specified that the FDR/PEP calculations should be consistent, with a definition of FDR associated with a probability score threshold s as

$$FDR(s) = \frac{\sum_{k=1}^{K} 1\{p_k \geq s\}(1 - p_k)}{\sum_{k=1}^{K} 1\{p_k \geq s\}}$$

where p_k is the probability of the presence of a proteoform; k is computed for $k = 1, \ldots, K$; and $1\{p_k \geq s\}$ is an indicator variable that takes on the value 1 when $p_k \geq s$ and 0 otherwise. The sum in the denominator is equal to the number of accepted proteoforms above the threshold s.

Participants were asked to describe the strategy they used for identification of the peptides/proteoforms and estimation of their FDRs. This description could be supplied as an (executable) R Markdown or IPython Notebook in which the narrative description was combined with R/Python scripts and results. Example R Markdown and IPython Notebook formats were provided on the study website. Alternatively, the method could still be described as free text in a plain text file, as in previous iPRG studies. Although the Markdown/Notebook formats were not required, it was hoped that participants would take advantage of these to create a more transparent and open study that permits sharing of executable scripts, as well as their results. The output generated by the example Markdown and Notebook already conformed to the specified results submission format. Furthermore, web-based hosting services, such as GitHub, directly support the viewing of R Markdown and IPython Notebook (.ipynb) in modern web browsers.

In previous studies, either a single link to a temporary file transfer server (*e.g.*, a file transfer protocol server) or a combination of a static web page and a file-transfer server was used for information distribution and participant submissions. It was often time consuming for the study organizers to set up and administer a physical server (*e.g.*, server configuration, software installation, and network firewall management) for hosting a study within a university, academic hospital, or a national laboratory. To address these problems, we proposed a lightweight and inexpensive web-based platform for the study, runnable on any compatible hosting service, including cloud servers and VPS, and manageable by free, open-source software. We used a VPS for this study, which significantly lowered the barrier for creating a website and allowed everyone in the iPRG to manage this virtual server remotely with less effort than in the past. We configured the Apache web server to run a back-end script, coded in PHP, with the Semantic UI and jQuery JavaScript libraries for the front end.

Issues detected by the submission validator were reported to the participant during the submission process. Technically, the submission validator was written in Python and was independently executed as an external program by PHP when users submitted their results through the web page. This validator raised exceptions when the submission files did not meet the requirements—especially for file headers, data types, and value ranges for rows and columns. The validator can be fully customized, according to the requirements of the study. Upon successful validation of the results and uploading of a method description file (in any of the approved formats), the submission was accepted and committed to a GitHub repository dedicated to this study. For anonymity, a random identifier was automatically generated by the submission system and communicated with the participant, unless one was provided by the participants themselves. This identifier could then be used by the participants to update or modify their own submission, while GitHub automatically traced all changes with version control. No private information potentially identifying the participants, such as e-mail or internet protocol address, was collected in the GitHub repository. After the study was completed, the repository was made public, allowing the participants (and others) to examine all of the methods and results (presented anonymously) and reproduce them. As the GitHub repository is independent from the VPS used for the submission system, it will be available even after the study site is taken down.

RESULTS

The iPRG 2016 study produced 12 high-quality liquid chromatography-MS/MS datasets with known proteoforms present, detailed and executable example R Markdown and IPython Notebooks containing entire proteomics data analysis pipelines for proteoform identification, and a VPS web-server template for future ABRF studies, which is available to all ABRF research groups or anyone else who wishes to set up a study similarly in the future.

The study ran until January 16, 2017, and resulted in 8 unique submissions. From the study survey, as well as the small number of participants, it is clear this was a difficult task, with 1 participant reporting having spent >10 h on the submission. The participants used 4 principal peptide identification workflows: 1) OpenMS[11] with Mascot,[12] MS-GF+,[13] and X!Tandem[14]; 2) Pyteomics[15] with DeMix[16] and X!Tandem and MP Score[17]; 3) MaxQuant[18] with Andromeda[19]; and 4) Trans-Proteomic Pipeline[20] with

X!Tandem and/or Comet.[21] Two participants only uploaded incomplete submissions. The submissions, including the answer key (which proteoforms were present in which sample), are available on iPRG2016.org and in the GitHub repository (www.github.com/iPRG-ABRF/2016_Study_ProteinInference).

Use of the format validator ensured that submissions could be directly analyzed and compared without the need for manual checking and correction of the submitted tables. In general, the submissions agreed on which proteoforms were present and which were not, and most methods assigned either a very low or very high probability of presence to any given proteoform. This high correlation is illustrated in the density plots in **Fig. 2**, showing a comparison between two submissions (24601 and cSQeHtHi). Some submissions (nXlg8ObD, pS01qvzb, and WDIpDUIQ) consistently overestimated the FDR, whereas 2 (24601 and cSQeHtHi) underestimated it (**Fig. 3**). On average, the four submissions reporting PEPs were accurate in their proteoform FDR estimates for the B and C datasets (the individual PrEST pools) and slightly conservative for the A datasets containing both pools (**Fig. 4**). The latter is not surprising, as all synthesized PrESTs were present in the sample, and all of these that claimed to be present are a true positive.

The supported R Markdown and IPython Notebooks make submitted methods reusable and submitted results reproducible in the sense that anyone able to run R or Python can repeat the analyses and regenerate the results, as well as apply and compare methods on other data than those used in the study. The scripts embedded in the example Markdown and Notebook helped precisely specify the expected submission format. Tools, such as RStudio or Jupyter, can also be used to generate effective, professional-looking reports in Hypertext Markup Language or Portable Document Format. Unfortunately, no participant shared his or her FDR calculation scripts, as shown in the provided examples. This was not entirely related to level of ability, as several participants reported having used in-house-developed R or Python scripts in the final step of their analysis. There was 1 accepted submission with the method description in an R Markdown, but Markdown was mostly used to format the text and produce a document with a clear layout; the R script was only used for plotting a heatmap of the results near the end. Although this was a creative use of the R script that was allowed in this study, it did not reflect the original intention of supporting R Markdown.

As mentioned above, the use of a VPS afforded several advantages for the iPRG. It was found to be easy to set up and adapted for the study, including running custom PHP and Python scripts. The hosting of the submission server outside of corporate/institutional firewalls was also a practical necessity, as a result of the custom submission validator scripts and automatic deposition of accepted submissions to the GitHub repository. There was also no interference with the ABRF website, as the virtual

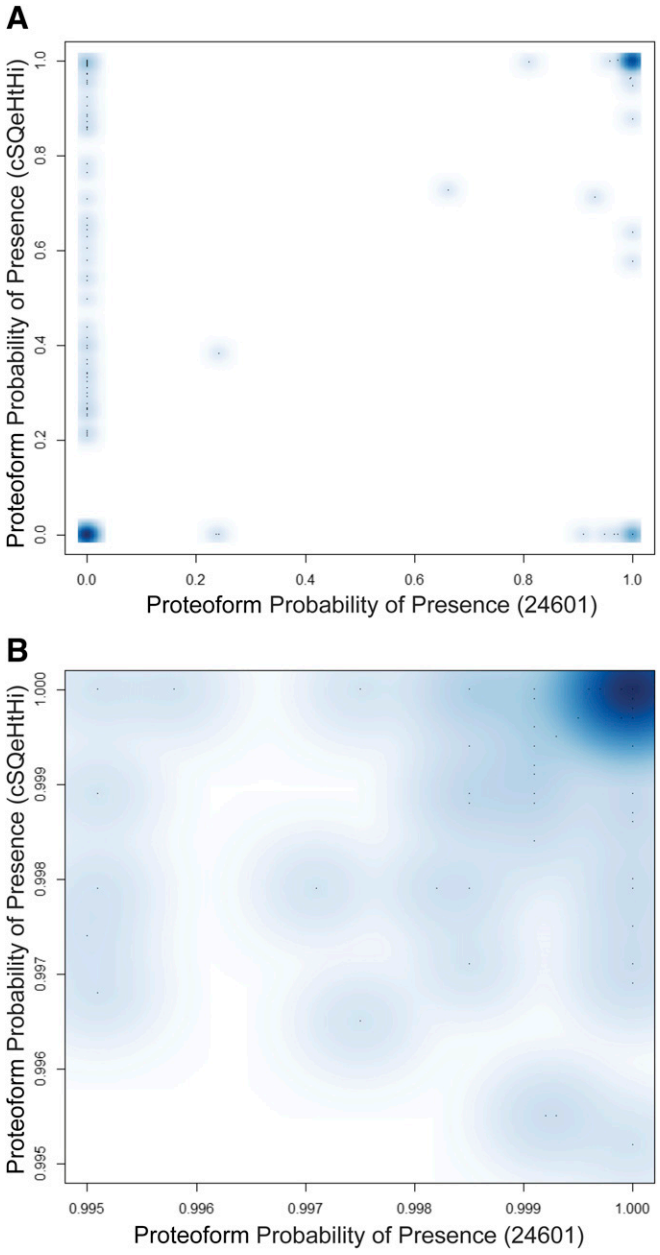

FIGURE 2

Comparison of 2 submissions for 1 dataset, showing a general agreement on which proteoforms are present, with probabilities near 1, and which are not, having probabilities near 0 (A). The color intensity represents the density of proteoforms at given probabilities of presence. However, some proteoforms are identified in one submission but not the other. There are also minor differences in the precise probabilities for confidently identified proteoforms (B).

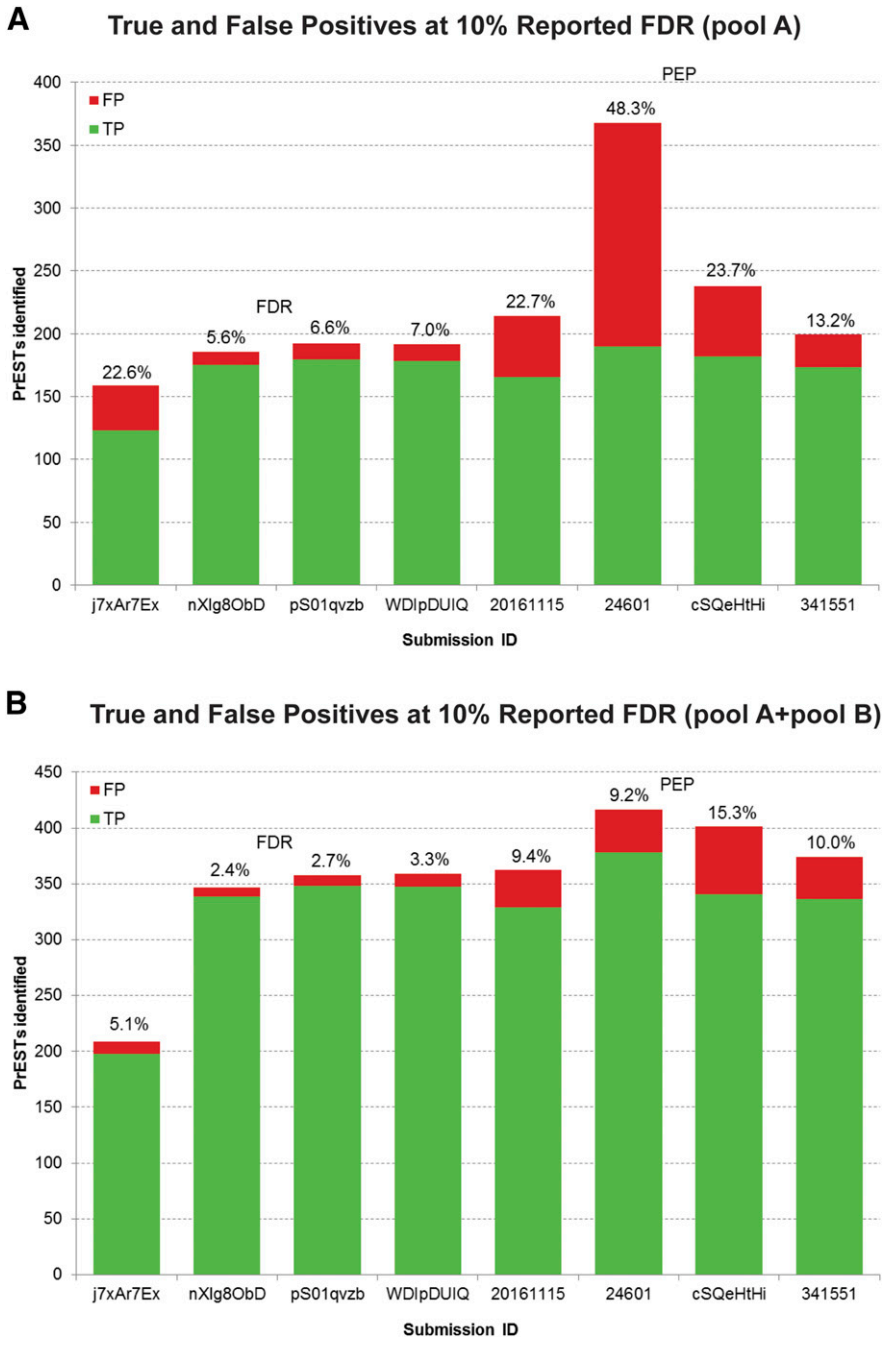

FIGURE 3

Participant-estimated and actual FDR at 10% estimated FDR for PrEST pool A samples (192 PrESTs, (A) and PrEST pool A and pool B samples (383 PrESTs; B). FP, false positive; TP, true positive.

machine resided on an entirely different server. Finally, it was inexpensive, especially compared with the time and effort that would have been required to accomplish the same results behind one of our institutional firewalls or using dedicated server hardware.

DISCUSSION

In summary, we learned that protein inference and proteoform-level FDR calculation are still challenging tasks. Most participants did quite well and erred on the conservative side, whereas sample A (the combined PrESTs) was practically a "best case" scenario. No one submission performed the best on all samples, so no gold standard can yet be established. The ground-truth datasets, generated specifically for the 2016 iPRG study, are unique and available to the research community for checking their protein inference methods, in particular, when dealing with homologous proteins.

In a recent European life sciences infrastructure for biological information meeting in Tübingen, metadata standardization and annotation of public datasets were identified as prioritized areas for the future of proteomics.[22]

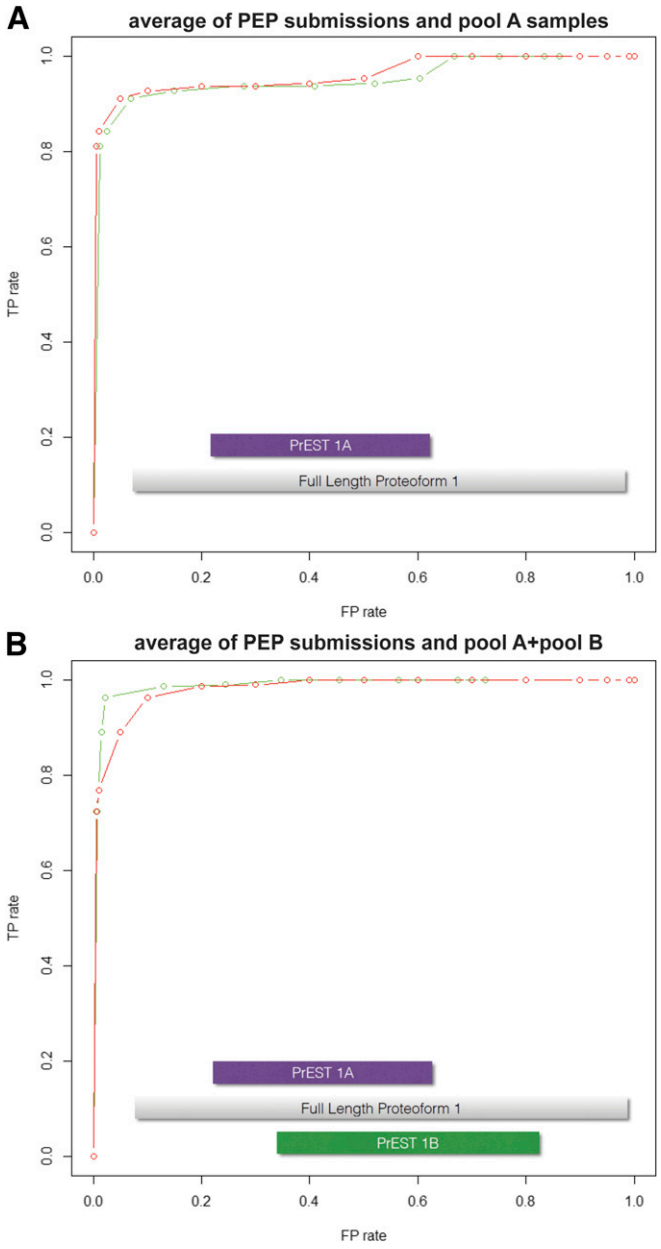

FIGURE 4

Average estimated FDR for the 4 PEP submissions (red) and actual FDR (green). The submissions were reasonably accurate in their FDR estimations of the PrEST pool A samples (A). For the datasets from the combined pools, where all expressed PrESTs were present, the methods were overly conservative at FDRs below 0.2 (B).

From the current study, we learned that there is still a gap between the general consensus on the importance and value of sharing metadata, such as methods, and the ability or willingness to do this in practice, even in a situation where one of the specified goals of the study was the sharing of methods. Although we provided example solutions to the study task as an R Markdown and an IPython Notebook, there is still a need for education in documentation and sharing of methods in bioinformatics.

The new submission mechanics and formats were evaluated as part of the study with the intention to make these available eventually to all ABRF Research Groups. Both the VPS web server and format validator worked well. None of the participants reported any problems, and no questions specifically related to the submission mechanics or file formats were asked *via* the e-mail address dedicated to the study (questions@iprg2016.org). The analysis and comparisons of the submissions were greatly facilitated by the format validator. With the number of submissions in this study, the validator was not absolutely necessary, but in larger studies or studies collecting multiple files in each submission, use of the VPS web server and format validator will be greatly beneficial. The PHP and Python scripts that were run on the website are available from the corresponding author.

ACKNOWLEDGMENTS

The authors thank Drs. Fredrik Edfors and Björn Forsström for generating data used in this study and acknowledge financial support from ABRF. S.H.P., R.W., and J.-Y.L. were supported by an Early Career Award from the Office of Biological and Environmental Research, U.S. Department of Energy. M.R.H. was supported by the National Institute for General Medical Sciences (Grant R01 GM087221), National Centers for Systems Biology (Grant 2P50 GM076547), and U.S. National Institutes of Health, National Heart, Lung, and Blood Institute (Grant R01 HL133135-01). L.K. was supported by a grant from the Swedish Research Council (2017-04030). The authors declare no conflicts of interest.

REFERENCES

1. Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* 2005; 4:1419–1440.
2. Audain E, Uszkoreit J, Sachsenberg T, et al. In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. *J Proteomics* 2017;150: 170–182.
3. The M, Tasnim A, Käll L. How to talk about protein-level false discovery rates in shotgun proteomics. *Proteomics* 2016;16: 2461–2469.
4. Choi M, Eren-Dogu ZF, Colangelo C, et al. ABRF Proteome Informatics Research Group (iPRG) 2015 Study: detection of differentially abundant proteins in label-free quantitative LC-MS/MS experiments. *J Proteome Res* 2017; 16:945–957.
5. R Studio. R Markdown. Available at: http://rmarkdown.rstudio.com/. Accessed January 8, 2018.
6. Loraine AE, Blakley IC, Jagadeesan S, Harper J, Miller G, Firon N. Analysis and visualization of RNA-seq expression data using RStudio, Bioconductor, and Integrated Genome Browser. *Methods Mol Biol* 2015;1284:481–501.
7. Grüning BA, Rasche E, Rebolledo-Jaramillo B, et al. Jupyter and Galaxy: easing entry barriers into complex data analyses for biomedical researchers. *PLOS Comput Biol* 2017;13: e1005425.
8. De Raad M, de Rond T, Rübel O, Keasling JD, Northen TR, Bowen BP. OpenMSI Arrayed Analysis Toolkit: analyzing

spatially defined samples using mass spectrometry imaging. *Anal Chem* 2017;89:5818–5823.
9. Uhlén M, Fagerberg L, Hallström BM, et al. Proteomics. Tissue-based map of the human proteome. *Science* 2015;347: 1260419.
10. The M, Edfors F, Perez-Riverol Y, et al. A protein standard that emulates homology for the characterization of protein inference algorithms. *J Proteome Res* 2018;17:1879–1886.
11. Pfeuffer J, Sachsenberg T, Alka O, et al. OpenMS - a platform for reproducible analysis of mass spectrometry data. *J Biotechnol* 2017;261:142–148.
12. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;20:3551–3567.
13. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* 2014;5:5277.
14. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004;20:1466–1467.
15. Goloborodko AA, Levitsky LI, Ivanov MV, Gorshkov MV. Pyteomics–a Python framework for exploratory data analysis and rapid software prototyping in proteomics. *J Am Soc Mass Spectrom* 2013;24:301–304.
16. Ahn J, Yuan Y, Parmigiani G, et al. DeMix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics* 2013;29:1865–1871.
17. Ivanov MV, Levitsky LI, Lobas AA, et al. Empirical multidimensional space for scoring peptide spectrum matches in shotgun proteomics. *J Proteome Res* 2014;13: 1911–1920.
18. Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* 2016;11:2301–2319.
19. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* 2011;10: 1794–1805.
20. Deutsch EW, Mendoza L, Shteynberg D, Slagel J, Sun Z, Moritz RL. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clin Appl* 2015;9:745–754.
21. Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. *Proteomics* 2013;13: 22–24.
22. Vizcaíno JA, Walzer M, Jiménez RC, et al. A community proposal to integrate proteomics activities in ELIXIR. *F1000 Res* 2017;6:875.